BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

PROTEINS: Structure, Function, and Genetics 7:317–334 (1990)

Comparative Modeling Methods: Application to the Family of the Mammalian Serine Proteases

Jonathan Greer
Computer Assisted Molecular Design Group, Pharmaceutical Products Division, Abbott Laboratories, Abbott Park, Illinois 60064

ABSTRACT Comparative modeling methods are described that can be used to construct a three-dimensional model structure of a new protein from knowledge of its sequence and of the experimental structures and sequences of other members of its homology family. The methods are illustrated with the mammalian serine protease family, for which seven experimental structures have been reported in the literature, and the sequences for over 35 different protein members of the family are available. The strategy for modeling these proteins is presented, and criteria are developed for determining and assigning the reliability of the modeled structure. Criteria are described that are specially designed to help detect cases in which it is likely that the local structure diverges significantly from the usual conformation of the family.

Key words: protein structure, computer modeling, structure prediction, sequence homology, structure homology

INTRODUCTION

It has been apparent for close to two decades that proteins from very different sources and sometimes with rather diverse functions can have homologous sequences and consequently very similar three-dimensional structures.[1] This fact has been the basis for the development of comparative modeling methods,[2–7] which permit extrapolation from the experimentally determined structure for one or more members of a homologous family to a new member of this family whose sequence has been determined but whose structure is as yet unknown. A number of factors combine to increase greatly the application of comparative modeling techniques today. The large number of protein structures[8] and the exploding number of protein sequences[9] that are being reported in the literature and that are readily available in computerized databases provide the basic structural and sequence data needed to apply the method. The proteins in which we are interested are often available in only small quantities, too little for structural studies unless the gene or mRNA is cloned or synthesized and expressed. Even when this latter effort is deemed worthwhile and is initiated, the comparative modeling studies can be performed in the meantime, providing an approximate view of the structure of the molecule until sufficient protein can be obtained and the experimental structure can be determined. Such a model structure can be very useful in the interim to help plan and interpret biochemical[10] and mutagenesis[11] experiments, to probe functional properties,[12] and improve our understanding of ligand or substrate binding.[13–15]

The use of comparative methods involves extrapolation from one or more known structures to construct a new structure. It is important to consider how accurately this function can be performed in assessing whether it is a useful and worthwhile exercise. Certainly, comparisons that have been published between predicted model structures and the experimental structures determined later[16–18] indicate that the modeled structure is not likely to be completely accurate. This paper reports the methods that we have developed to perform comparative modeling and the criteria that we use to estimate the reliability of various parts of the structure and especially to identify potential problem areas that are likely to be particularly difficult to predict correctly.

To describe properly and illustrate the modeling techniques, we will use the mammalian serine protease family.[19] Members of this family are ubiquitous in nature; they are present all along the evolutionary pathway from bacteria to humans. They play important roles in a wide variety of body functions, including blood coagulation, fibrinolysis, complement activation, fertilization, and digestion. Therefore, modeling the various members of this family can provide valuable new information about divers critical biological functions of the body.

Comparative modeling works best when there are several experimental structures and known sequences. Experimental structures for seven differ-

Received September 12, 1989; accepted January 17, 1990.
Address reprint requests to Dr. Jonathan Greer, D-47E, Pharmaceutical Products Division, Abbott Laboratories, 1 Abbott Park Road, Abbott Park, IL 60064.

TABLE I. Experimentally Known Three-Dimensional Structures of Serine Proteases

Protein	Source	Resolution (Å)	Reference
Chymotrypsin	Bovine	1.8	24*
Trypsin	Bovine	1.8	25*
Elastase	Porcine	1.8	26
Kallikrein	Porcine	1.8	27
Mast cell protease	Rat	1.8	28
S. griseus trypsin	*S. griseus*	1.7	23
Tonin	Rat	1.8	29

*Structures have been determined by several groups for each of these proteins. The reference shown corresponds to the coordinates used in this work.

ent serine proteases* that can be considered members of this mammalian family have been reported in the literature and deposited in the Brookhaven Protein Structure Database.[8] These include chymotrypsin,[24] trypsin,[25] elastase,[26] kallikrein,[27] rat mast cell,[28] *Streptomyces griseus* trypsin-like protein,[23] and tonin[29] (see Table I). Several of these have been solved independently in more than one laboratory either in the same form or in alternate forms of the molecule.[30-32] In addition, structures have been reported for the precursor, zymogen, form of chymotrypsin[33,34] and trypsin.[35,36] To complement these structures, amino acid sequences for over 35 different serine proteases have been reported (Table II), not counting species variations. Thus this family presents an ideal system for developing comparative modeling methods and at the same time applying them to proteins involved in important and interesting biological functions.

COMPARATIVE MODELING METHOD: THE "SPARE PARTS" ALGORITHM

Comparative modeling requires extrapolation from known structures to produce a model of an unknown structure. Our ability to extrapolate accurately is, unfortunately, greatly limited by our still rudimentary knowledge and understanding of protein structure and energetics. Consequently, the techniques that are used to extrapolate influence the resultant model critically and may introduce considerable error. The methods that we have developed attempt to systematize the modeling process by combining all the known structures and sequences to help improve the accuracy of the extrapolation.

*Structures have also been reported for several serine proteases from bacterial sources.[20-22] Although these are certainly members of this family, they differ sufficiently in their structures that comparative modeling methods are not developed enough to permit modeling them from the mammalian structures without serious errors.[23] Therefore, they are not considered in this work.

Analysis of the Structural Properties of the Family

The first steps in the modeling process are illustrated schematically in Figure 1. Let us assume that we have experimentally determined three-dimensional structures for several members of the homologous family of interest, e.g., structures "A," "B," and "C" in Figure 1a. The known structures are superimposed in three dimensions to obtain a maximal overlap of the structures (Fig. 1b). Performing this superposition of the structures is sometimes not a trivial procedure. This is true because we want to overlap the molecules based on the parts that are the same and ignore the parts that are different. The more different the molecules are from each other, the harder it is to find a unique best overlap of the common features of the structures. Several programs have been written that do this analytically.[37,38]

Once superimposed, there are parts of the known structures that overlap very well, indicating that the structures are closely conserved in these regions (see bold areas in Fig. 1b). We call these portions "structurally conserved regions," or SCRs, and expect that they will usually remain conserved in all members of this homologous family. These regions are usually composed of the secondary structure elements, the immediate active site, and other essential structural framework residues of the molecule. Between these conserved elements are highly variable stretches that differ significantly from one member of the family to the next. These are called "variable regions," or VRs. They are almost always loops that lie on the external surface of the protein, and they contain all the additions and deletions between different protein sequences.

After the structures have been overlapped in three dimensions and parsed into SCRs and VRs, the next step is to align their amino acid sequences. For purposes of comparative modeling, the sequence alignment is done differently from the usual methods.[39] Instead of relying on criteria such as amino acid identity or homology, we use strictly the three-dimensional overlap as the criterion. When the α-carbons of the respective residues in the overlapped protein structures occupy the same place in three-dimensional space, then the residues are corresponded in the sequence alignment (Fig. 2). Thus the alignment is primarily concerned with the SCRs, since these are the portions of the structures that are basically the same in all protein members of the family. The alignment in the VRs is often arbitrary, unless two or more structures have similar conformations in a particular VR (for an example of this see the VR in the upper left of the structures "A" (sequence C-D-F-A) and "C" (sequence C-R-Y-V) in Figs. 1 and 2).

The resultant sequence alignment is then scruti-

TABLE II. Sequences of the Serine Proteases*

Protein	Code	Source
Chymotrypsin	CHT†	Bovine
Trypsin	TRP†	Bovine
Elastase	ELA†	Porcine
Kallikrein	KAL†	Porcine
Mast cell protease	MCP†	Rat
S. griseus trypsin	SGT†	*S. griseus*
Tonin	TON†	Rat
Haptoglobin heavy chain	HPH	Human
Protein Z	PRZ	Bovine
Protein C	PRC	Human
Nerve growth factor α chain	NGA	Human
Nerve growth factor γ chain	NGG	Human
Blood clotting factor VII	VII	Human
Blood clotting factor IX	FIX	Human
Blood clotting factor X	FAX	Human
Blood clotting factor XI	FXI	Human
Blood clotting factor XII	XII	Human
Plasmin	PLM	Human
Apolipoprotein A	ALP	Human
Tissue plasminogen activator	TPA	Human
Urokinase	UKH	Human
Thrombin	THR	Human
Complement factor B	CFB	Human
Complement factor 2	CF2	Human
Complement factor D	CFD	Human
Complement factor 1R	C1R	Human
Complement factor 1S	C1S	Human
Adipocyte protease	ASP	Mouse
Cathepsin G	CAG	Human
T-cell serine protease	TCL	Mouse
Hannuka factor	HAF	Human
Cytotoxic T lymphocyte protease	CTL	Mouse
Batroxobin	BTX	Snake
EGF binding protein	EBP	Mouse
Drosophila snake locus	DSN	*Drosophila*

*The sequences are taken from the sequence database.[8]
†For those proteins for which a three-dimensional structure is available, the sequence is taken from the Brookhaven structure database.[9]

nized carefully to identify strong stretches or patterns of sequence homology that are characteristic of each SCR in the structure. For example, the sequence "L-S/T-V-π-I-π," where π is a charged or polar residue, is conserved in the second SCR in Figure 2. Similarly, "G-I-A" is found in all members at the third SCR. However, in the fourth SCR, the situation is less clear. A proline usually appears in the second position but may be replaced by a glycine. Most SCRs can be identified by a characteristic (not necessarily contiguous) sequence pattern. In some cases, there is a pattern of hydrophobic or hydrophilic residues rather than specific side chains. Careful analysis of the sequences and the corresponding three-dimensional structures will usually allow some pattern to be discerned.

The identification of these sequence homology patterns is a crucial step in the modeling. As is shown below, it is essential for the correct alignment of a "new" sequence. If the characteristic sequence pattern is not present, then it is not clear how to align the "new" sequence and, consequently, how to construct the model structure. This is one of the reasons that the comparative modeling method currently depends so heavily on the retention of sequence homology. The above steps are performed once for a protein family and need be reexamined only as new experimental structures and their sequences become available.

Construction of a "New" Structure

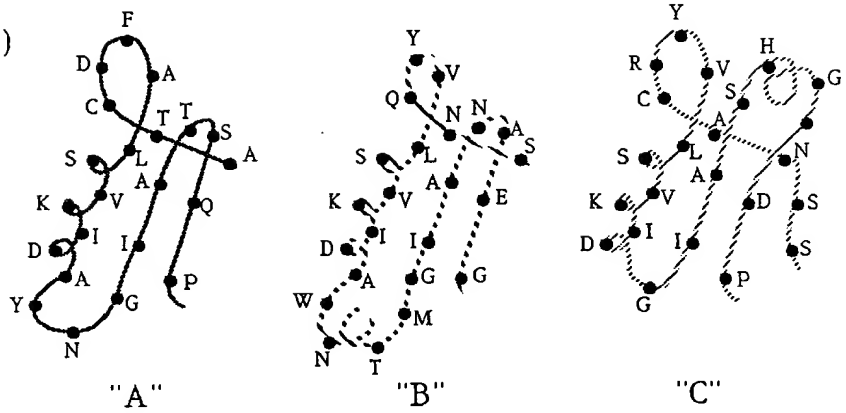
We are now ready to begin modeling a "new" protein sequence of interest that is a clear member of this homologous family. The first step is to align this "new" sequence to the sequences of the SCRs using the previously identified characteristic sequence homology patterns. It should be possible to align an appropriate portion of the "new" sequence to each SCR with these patterns. Thus, the "new" sequence L-T-V-K-I-E fits the pattern L-S/T-V-π-I-π described above and can be aligned to the SCR at positions 7–12 in Figure 2 and similarly for the G-I-A sequence of the SCR at 16–18. There will occasionally be more than one possible alignment for a particular SCR, especially if the sequence pattern is a weak one. Such an example might occur in the SCR corresponding to residues 1–2 (Fig. 2). In such cases, the different possible alignments would have to be considered. As each portion of the "new" sequence is aligned to an SCR, the rest of the positions in that SCR are filled with the adjacent "new" sequence without permitting any additions or deletions within the SCR.†

The remaining residues make up the VRs. The "new" sequence in each VR is examined to see if it corresponds in length and residue character to one of the VRs of the known structures. For example, the third VR in the "new" protein in Figures 1 and 2, residues 13–15, has the same sequence length and character as this VR in protein "B" (A-W-D-S-L in "new" vs. A-W-N-T-M in "B") and therefore has been aligned to it. On the other hand, none of the known structures has a sequence that corresponds to residues 3–6 of the "new" protein in residue length.

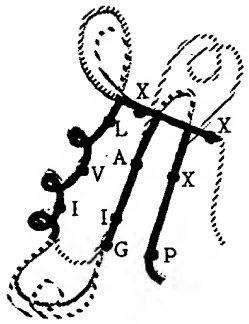
Once the "new" sequence is aligned, as shown in Figure 2, the model building can begin. For the SCRs, usually the main chain coordinates from any one of the known structures can be taken (Fig. 1c). The side chains are mutated to those of the "new" sequence wherever necessary. In our implementation, the χ_1 angle for the "new" side chain is chosen to maximize overlap of this side chain on the old one. Further side chain χ angles are not fitted at this

†An exception would occur if there are too few residues between this SCR and the neighboring SCR. In that case, the end residues of an SCR would be left with a deletion. For examples, see position 96 in PLM and ALP in Figure 4.

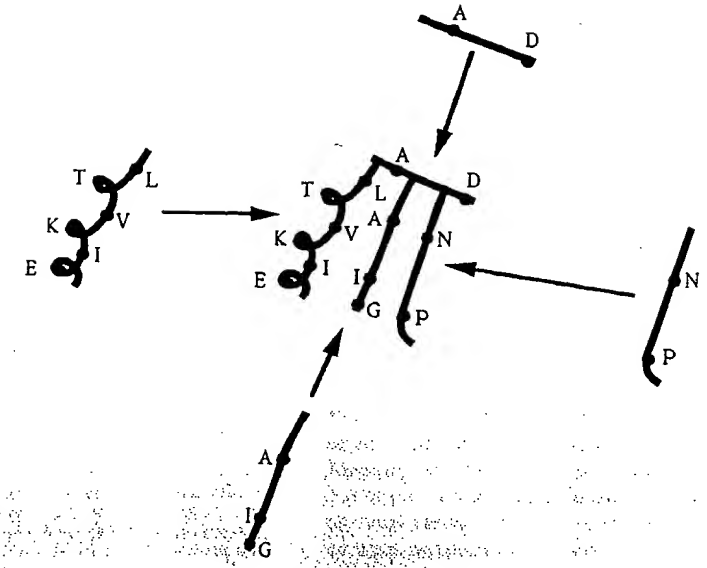
(a)



(b)



(c)



Figs. 1a–c. Legend appears on page 321.

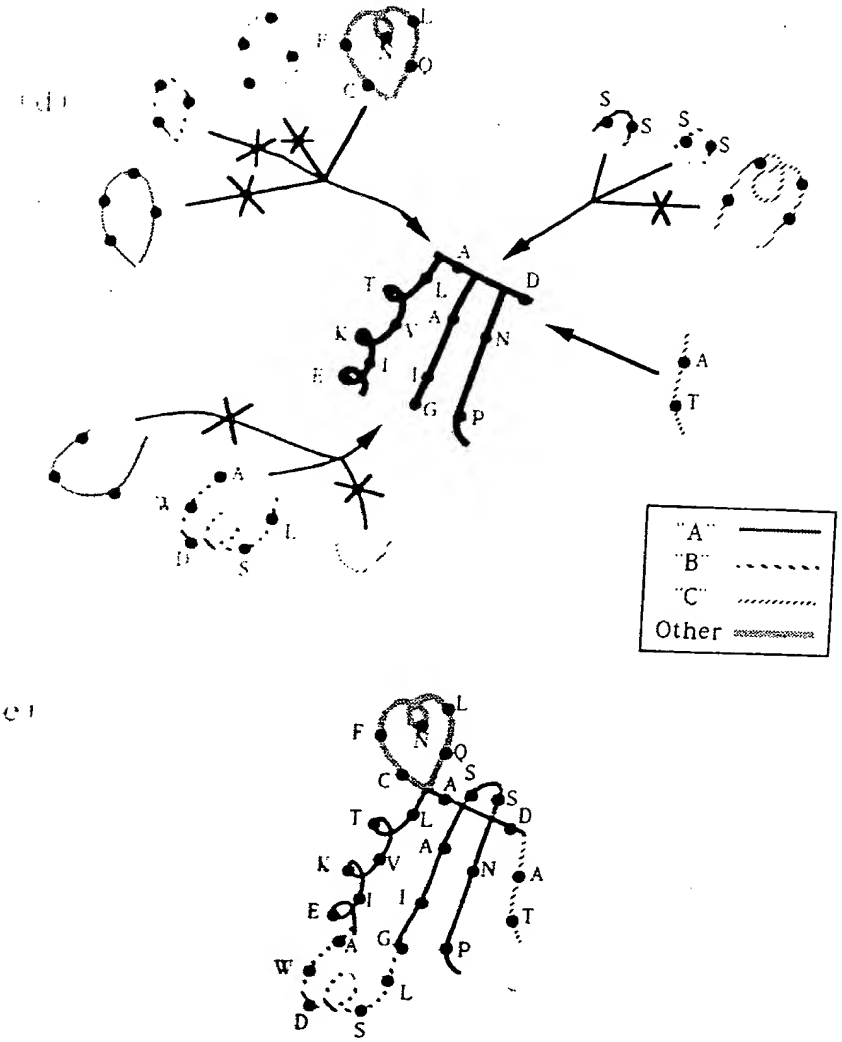


Fig. 1 Schematic representation of the comparative modeling method. **a:** There are three proteins in this homologous family whose structures are known, A, B, and C. Each protein is represented with a characteristic dashed or dotted pattern throughout the figure (see key in d). **b:** The three proteins are superimposed showing that parts of the structure are conserved (SCRs in bold lines) and parts are variable from one protein to the next (VRs, respective dashed and dotted lines). **c:** Steps in the construction of the schematic "new" model structure are presented in this and the next two parts. The SCRs (bold lines) are constructed from the main chain coordinates of any one of the known structures, since they are all very close to the same. The side chains are mutated to the "new" sequence where necessary. **d:** The various VR conformations found in the known structures are considered for each VR of the "new" protein. (Compare the respective VRs shown here with those in a and b.) The ones that do not fit are rejected, as shown by the crossed arrows. The most suitable is selected in each case. In some cases, other conformational search[41–45] or energetics[46,47] methods must be employed, since no suitable conformation can be found for that VR among the known structures (see VR in upper left corner with the sequence C-F-N-L-Q for an example in which a different, new conformation is necessary). **e:** The composite structure showing the source of the respective "spare parts" selected for the model structure.

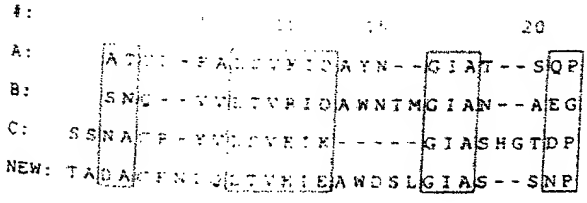


Fig. 2. Sequence alignment for the set of proteins in the schematic homologous family corresponding to Figure 1. The alignment is performed based solely on the overlap of the three-dimensional structures and not by sequence alignment methods. The boxes delineate the SCRs as determined from the three-dimensional structure overlap. The sequence of a "new" protein is aligned based on the characteristic patterns of sequence homology found for the known structures and their sequences (see text). The single letter amino acid codes are given in Figure 4.

stage, because the nature of the different side chains usually diverges beyond this point. An automatic angle-scanning energy-optimizing routine is usually not employed, because we find that it often selects an unsuitable conformation based on minor or insignificant differences in energy. However, incorporation of recently compiled and reported rotamer libraries[40] may improve our ability to automate the side chain conformation selection process.

For those VRs for which an appropriate example structure can be selected from among the known structures during the alignment process described above, the main chain fragment is taken directly from the VR of that known structure, and again the side chains are mutated to fit the "new" sequence (see examples in Fig. 1d and e). Remaining VRs, such as the VR at positions 3–6 in Figures 1 and 2, have no good example to build upon among the known structures. Therefore, they have to be constructed by more complex conformation search methods[41–45] and energetic considerations.[46,47] We have also used the protein database as a source of tentative starting conformations for loops of this type. The two ends of the particular loop that lie in the adjacent SCRs are taken, and the Brookhaven Protein Structure Database[8] is searched for structural fragments that closely match the conformation of these ends (specifically, the α-carbon positions) and have the appropriate number of residues between the ends. Thus, for example, in the VR at positions 3–6 of the "new" sequence, the structures of the two ends corresponding to residues "A-D-A" (positions −1 to 2) and "L-T-V- . . ." (positions 7 to 9. . .) are selected and typically ten structures that match these two ends best [lowest root mean square (rms) deviations] and that have five residues in between are chosen and examined. Obviously, any such conformation that would collide with the remainder of the constructed "new" structure is eliminated. Similarly, if the conformation does not pack well or buries a single charged group, it is rejected. In this way, one or a small number of initial conformations can be selected for this loop. Clearly this latter method does not produce an exhaustive list; however, it does provide loop conformations that are known to appear in other proteins. This method has previously been described by Kraulis and Jones[48] for constructing protein structures from fragments using nuclear magnetic resonance (NMR) data or using crystallographically generated electron density maps.

It is important to emphasize that it is the initial spatial alignment of the three-dimensional structures performed above (Fig. 1b) that allows this convenient clipping of fragments or "spare parts" from the various known structures to construct a composite model structure of the "new" protein. Thus overlapping of the structures is essential for two crucial steps: for the original alignment of the sequences and for clipping together fragments in the subsequent construction of each "new" model protein structure. Note that, the more experimentally known structures available, the more likely the boundaries of the SCRs will be well-defined and the greater the likelihood that a suitable example known structure, i.e., spare part, will be found for each "new" VR among the known structures.

Because the different portions of the "new" constructed model structure arise from quite different sources (Fig. 1e), we can assign appropriate qualitative reliability confidence levels to the respective parts of the structure. Clearly, the conformations in the SCRs are the most reliable, especially when several known structures are available and have been compared in detail. This is true only if the respective SCRs of the "new" sequence retain the characteristic homologous sequence patterns in those SCRs. When they do not, and this is illustrated below, it should be regarded as a red flag warning that something different may be happening in this region of the particular "new" protein. In such cases, great care must be taken in constructing the "new" structure and the confidence level in this portion of the molecule reduced appropriately. For the VRs, when a good model structure appears for the respective VR among the known structures, then a fair confidence level can be assigned. This level is not as high as in the SCRs because of the greater inherent variability of these loop regions. The actual confidence level for each such VR would depend on the size of the loop (larger loops have more degrees of freedom and are therefore less reliably determined), how good the sequence homology is to the known structure selected to model this VR, and how well the chosen conformation fits and packs onto the rest of the "new" structure. The confidence is lowest for those VRs with no good model loop among the known structures. When conformational methods or database search techniques have to be employed, experience[16–18] shows that we are not able to predict the conformation reliably in many of these cases.

The resultant structure is then examined in detail to identify possible serious errors. For example, the inner cores of the structure are checked to be sure that no inappropriate charges have been buried. The modeling method makes it unlikely that there are serious steric contacts between main chain atoms in the model. This is because the main chain coordinates in the SCRs were taken from the known structures. Only if the bad contact was present in the parent known structures will it be found in the model. In the VRs, absence of main chain overlap was a major criterion for acceptance of a conformation. Therefore, in practice, bad steric contacts can usually be relieved by side chain rotations; occasionally by having to select an alternative conformation for a VR. The model structure can then be introduced into an energy-minimization program[46,47] to

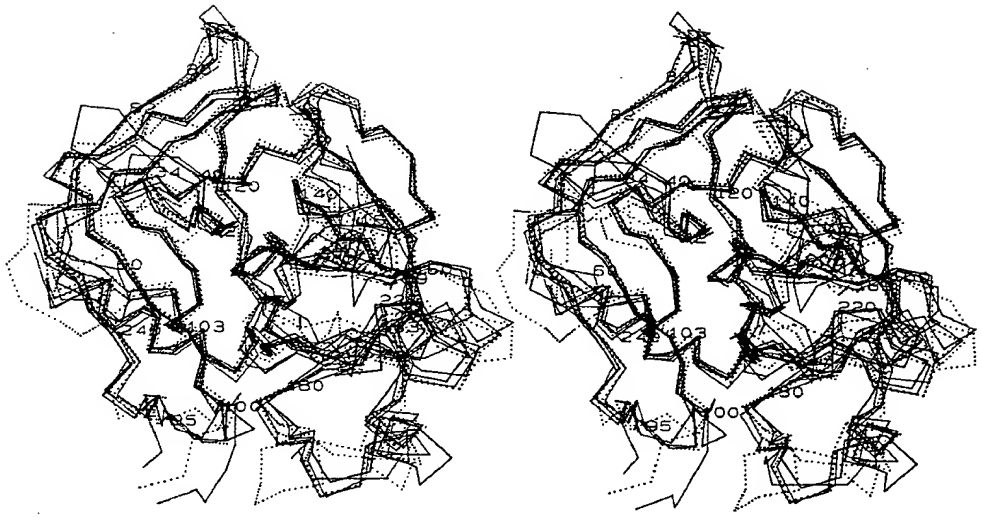


Fig. 3. α-Carbon plots showing the superposition of the seven experimentally known serine protease structures (see Table I). Note the SCRs, where the seven structures are virtually the same, and the VRs, where the structural divergence is considerable. The key to the plots is as follows: CHT, green; TRP, solid red ELA, solid blue; KAL, solid cyan; MCP, dotted green; TON, dotte red; SGT, dotted blue.

relieve any remaining minor steric contacts and to optimize the bond angles and torsional angles. Typically, we begin with restrained or template-forced[49] minimization, forcing the atomic coordinates to remain close to their initial positions. This allows bad contacts to relax without introducing large artifactual distortions into the structure. After several hundred cycles, the constraints can be progressively removed and the energy of the molecule minimized.

Experience with comparative molecular modeling[16–18] has shown that these models are never completely accurate, especially in the conformations selected for the VRs. Therefore, any uses of the model structure and predictions that are made based on it should take into consideration this fact and the local reliability confidence levels discussed above. We always regard such a model as a working hypothesis; as new data emerge, whether from structural, spectroscopic, or biochemical sources, the model is modified, corrected, improved, and refined to fit the new data.

RESULTS AND DISCUSSION

Initial Alignment of Known Serine Protease Structures

Our analysis of the serine proteases began some years ago with the superposition of the then three known structures of chymotrypsin, trypsin, and elastase.[3,4] Since then, coordinates for four more structures have appeared in the literature and data bank[8]: kallikrein,[27] mast cell protease,[28] *Streptomyces griseus* trypsin-like protease,[23] and tonin.[29] The overlap of these additional structures was performed by first superimposing the added experimental structure onto the rest using the α-carbon positions of residues 195, 57, and 102.‡ The superposition was refined, when necessary, using repeated cycles of least squares fit on related residue α-carbon positions between the protein to be fit and chymotrypsin. Only residues whose positions were conserved in the two structures were fitted in the calculation; that is, any residue was eliminated from the fitting if the deviation in α-carbon positions was more than 2 Å on the first cycle and then more than 1.5 Å on subsequent cycles. Figure 3 shows the final overlap obtained for these seven structures. Virtually all the SCR positions identified in the previous analysis of the first three proteins[4] remain structurally conserved. The repertoire of conformations, i.e. "spare parts," found for each of the VRs is increased with the larger number of structures, as expected (Table III).

‡The residue numbering given for the serine proteases follows that of the chymotrypsinogen molecule throughout this paper, in the text, in the tables, and in the figures.

TABLE III. Sequence Lengths of the Different VRs in the Serine Proteases

	VR									
VR	35–41	59–62	72–80	97–101	125–133	146–151	166–179	185–189	203–206	217–224
CHT	7	4	9	5	9	4	14	3	4	8
TRP	5	3	9	5	7	6	14	5	0	8
ELA	10	5	9	7	9	5	16	4	4	10
KAL	6	3	9	9	7	8	14	5	0	9
MCP	10	3	9	5	9	5	13	5	0	6
SGT	2	8	7	3	7	5	15	6	5	8
TON	4	3	9	16	7	6	14	5	0	9
HPH	6	16	0	1	8	4	31	5	6	7
PRZ	7	1	4	5	11	2	13	3	4	6
PRC	7	5	9	5	13	10	14	5	4	8
NGA	6	3	9	16	7	6	14	5	0	9
NGG	6	3	9	12	7	6	14	5	0	9
VII	6	8	9	5	12	5	19	5	4	8
FIX	6	5	9	7	11	5	14	5	4	8
FAX	7	5	8	5	11	5	14	5	4	8
FXI	9	8	9	5	9	5	15	5	3	8
XII	5	8	9	5	9	6	16	5	7	8
PLM	7	8	9	−1	9	4	16	5	4	8
ALP	7	8	9	−1	9	4	7	5	4	8
TPA	12	8	9	5	9	6	16	11	4	8
UKH	11	8	9	7	9	6	16	5	4	8
THR	8	13	10	6	12	11	14	8	6	8
CFB	10	8	1	14	25	3	27	7	4	14
CF2	6	8	5	14	25	0	28	4	4	21
CFD	6	7	10	5	11	5	10	3	0	9
C1R	4	12	8	8	9	3	19	5	6	6
C1S	3	5	10	12	11	4	23	4	7	5
ASP	6	8	9	5	11	5	16	3	0	9
CAG	9	3	9	5	9	4	14	5	0	6
TCL	6	4	8	5	9	5	18	5	0	9
HAF	6	4	9	5	9	5	18	5	0	10
CTL	9	3	9	5	9	5	15	5	0	6
BTX	6	3	13	2	7	6	13	5	0	9
EBP	6	3	9	16	7	6	14	5	0	9
DSN	15	5	10	5	6	5	20	4	7	8
Min.	2	1	0	−1	7	0	7	3	0	5
Max.	15	16	13	16	25	11	31	11	7	21
Range	13	15	13	17	18	11	24	8	7	16

Based on the above superposition of structures, the seven sequences were aligned as described above (Fig. 4). Note the wide variation in the length of the different sequences in each of the VRs (Table III). Using this sequence alignment, we attempted to identify the critical conserved stretches of amino acid sequence. These are characteristic of each SCR and are essential for aligning "new" sequences to the known ones, which is the first step in modeling. The patterns that were chosen are noted in Figure 4 at the bottom of the sequence list in the row labeled "CON." In some cases, there was almost complete conservation of a single amino acid. In other cases, only a hydrophobic side chain (typical of an internal position) seems to be required.

Because the identification of these stretches is such an important part of the modeling process, it is worthwhile describing in more detail how this is performed. In most cases, it is straightforward; the conserved sequence is readily apparent. For example, the characteristic I/V-I/V-G-G at positions 16–19 (see Fig. 4) defines the first SCR very clearly. Similarly, the remarkably conserved S/T/A-G-W-G sequence of residues 139–142 is highly characteristic of this SCR. On the other hand, the SCRs at 63–71 and 81–96 have no such clearly conserved sequence, and it is often very difficult to align a "new" sequence to the known ones in these SCRs unambiguously.

To understand better why this latter group of SCRs has so little sequence homology, the structures of these two regions were examined in detail (Fig. 5). It is immediately apparent that these SCRs lie on the surface of the molecule, and virtually all the residues point out into solvent and thus can and do vary with little consequent effect on the rest of the structure. However, in each SCR, there are some side chains that point into the interior of the molecule and thus are more conserved. In the SCR at 63–71, this includes the Val (or comparable aliphatic residue) at 66 and the aliphatic side chain at 68. The Gly at position 69 appears to be almost com-

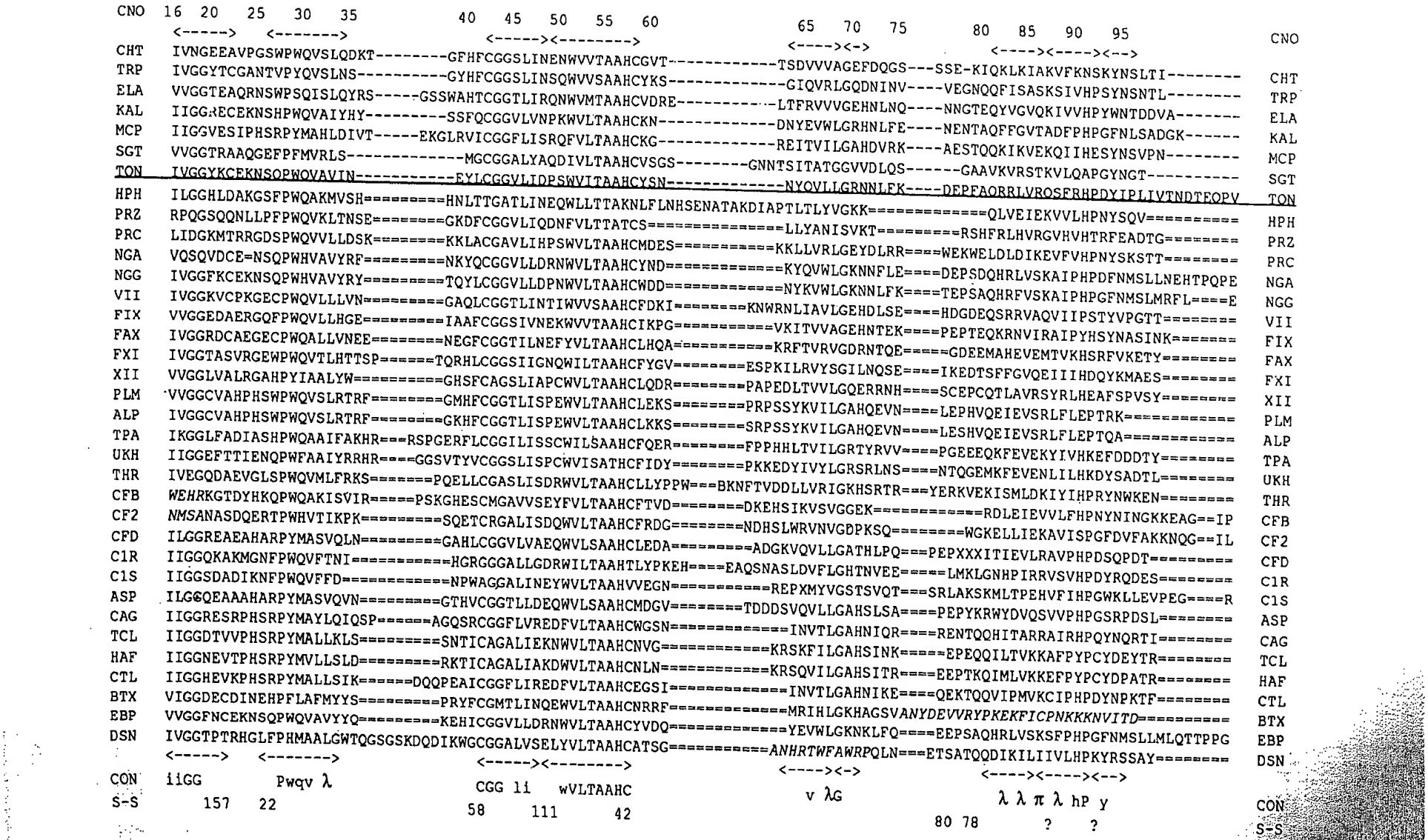


Fig. 4. Part 1. Legend appears on page 328.

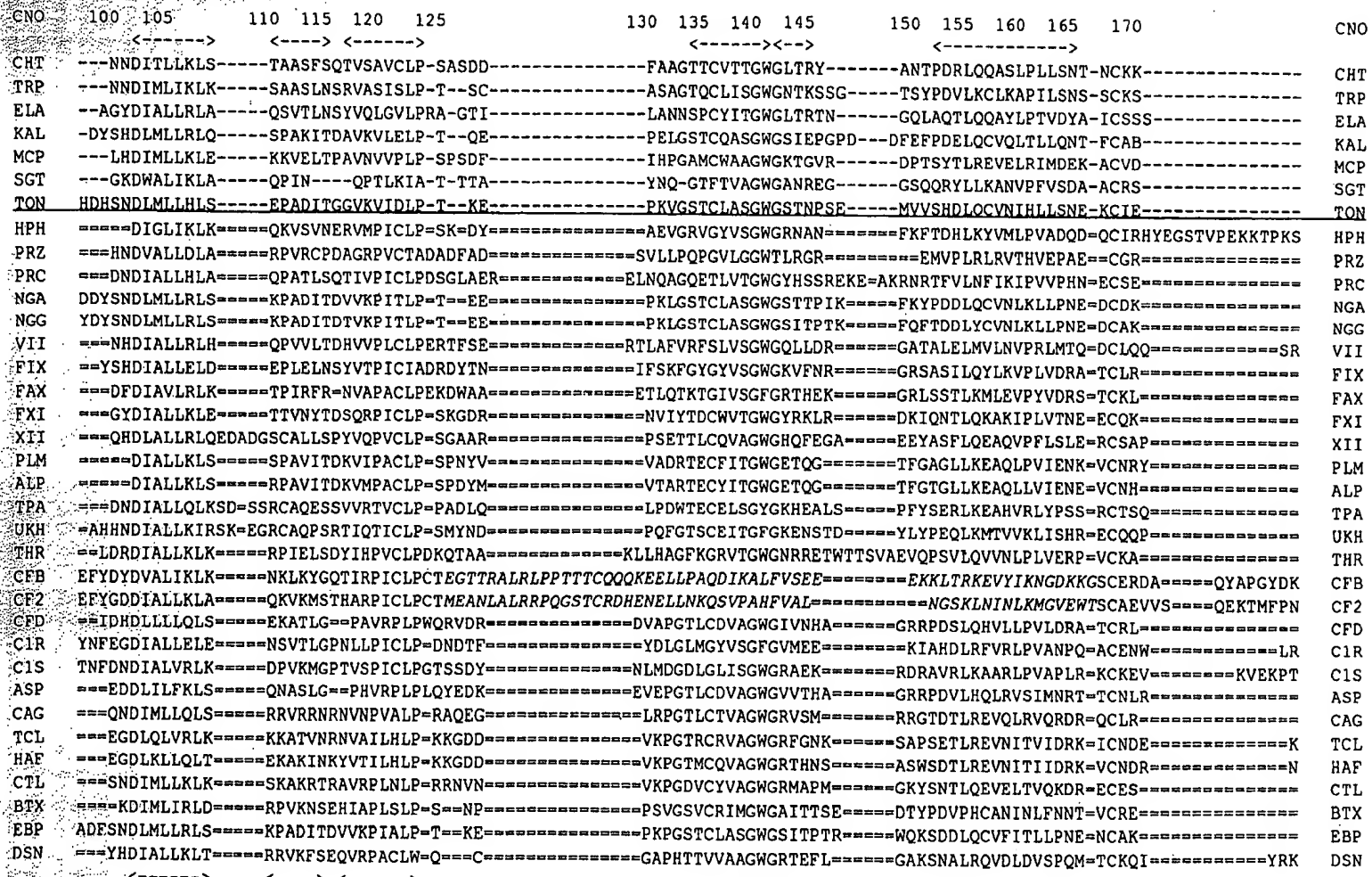


Fig. 4. Part 2. Legend appears on page 328.

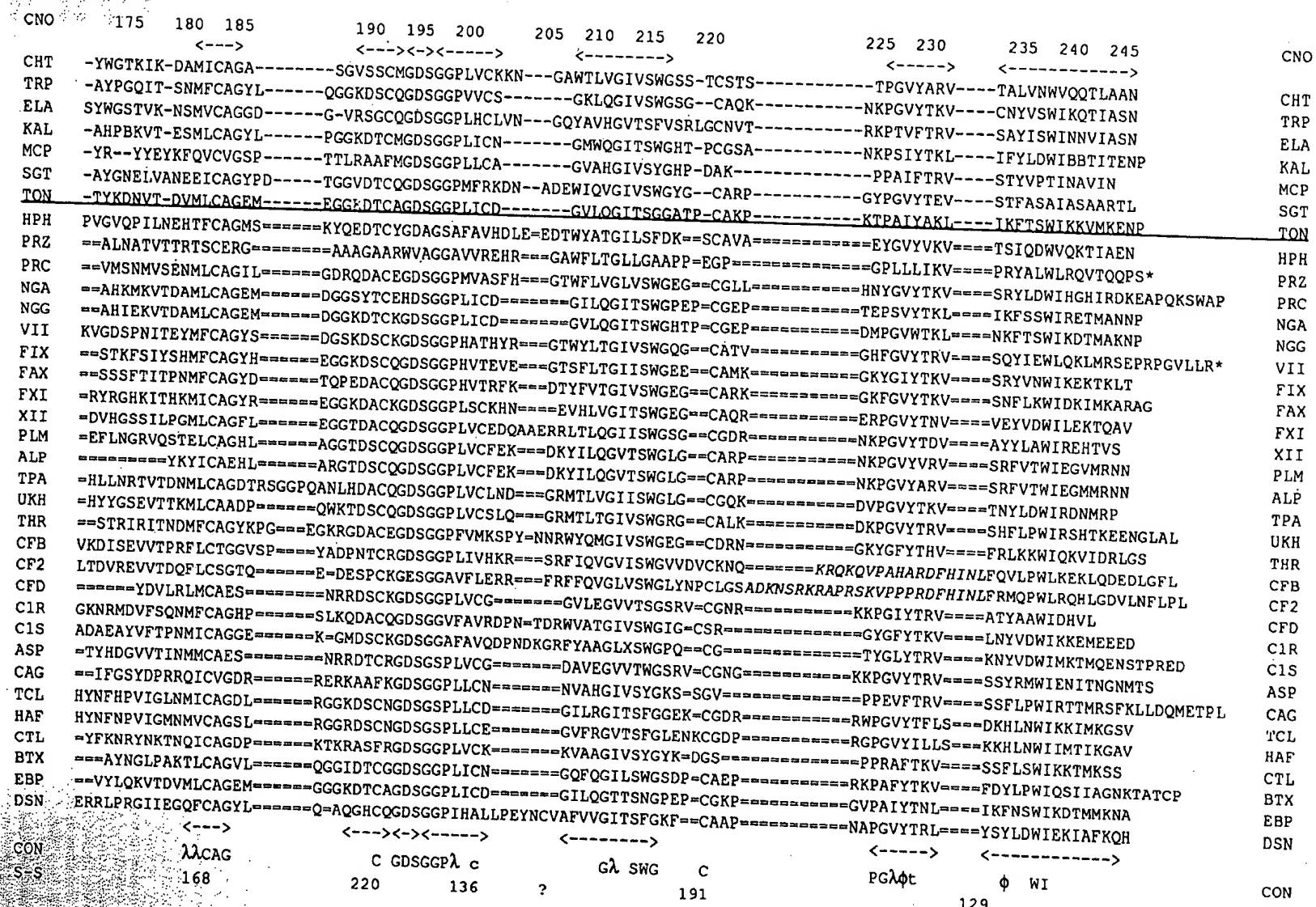


Fig. 4. Part 3. Legend appears on page 328.

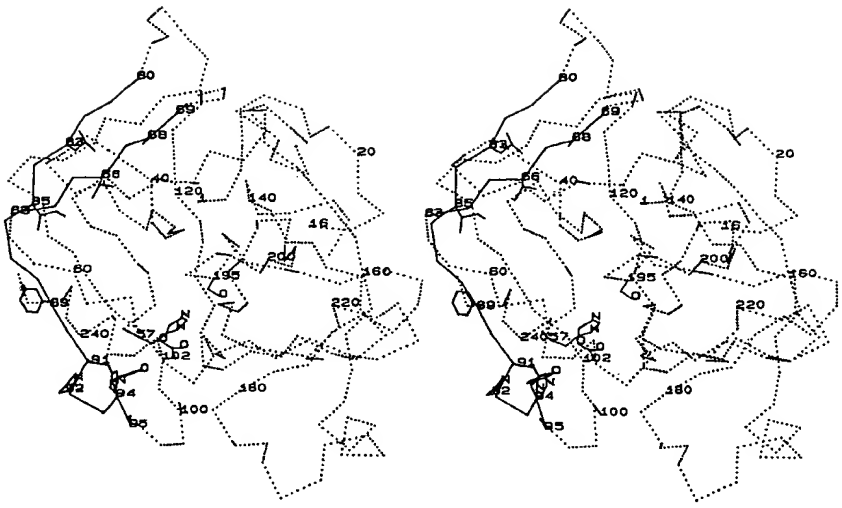


Fig. 5. The location is shown for some of the SCRs that are difficult to align in the serine proteases because of only weak characteristic patterns of homology. This is usually due to the SCR lying on the surface of the molecule. This figure presents the Cα plot of CHT (dotted lines), with the location shown of the respective SCRs that are difficult to align: 63–71 and 81–96 (solid lines). Note that these SCRs lie on the surface of the molecule and that most of their side chains point out into solvent and thus are free to vary more rapidly than internal residues. In the SCR at 63–71, conserved residues 66 and 68 are the only internal ones and are always hydrophobic; 69 is always a Gly, because it forms a turn. In the SCRs at 81–96, internal hydrophobic residues are 83, 85, 89, and 94. Partial conservation is observed at positions 91 and 92, with a His-Pro.

pletely conserved, probably because it forms a turn and adopts a conformation ($\phi = \sim 60°$, $\psi = \sim 25°$) that is permitted only for Gly (Fig. 5). Similarly, small hydrophobic side chains occur at positions 83 and 85 in the SCR at residues 81–86 (Fig. 5). More varied but always hydrophobic residues appear at 89, and an aromatic one is typical at 94, both of which point into the molecule. In addition, there is also a tendency for a His-Pro at positions 91–92 (Fig. 5); however, a number of sequences have only one or none of these two residues (Fig. 4).

```
CNO    55   60                  65   70   75    80        CNO
         --->                  <---->-<->      <-
CHT    AAHCGVT-------------TSDVVVAGEFDQGSSSE-KIQK        CHT
TRP    AAHCYKS-------------GIQVRLGQDNINV-VEGNQQF        TRP
ELA    AAHCVDRE-----------LTFRVVVGEHNLNQ-NNGTEQY        ELA
KAL    AAHCK--------------NDNYEVWLGRHNLFE-NENTAQF       KAL
MCP    AAHCKG-------------REITVILGAHDVRK-AESTQQK        MCP
SGT    AAHCVSGS--------GNNTSITATGGVVDLQS---GAAVK        SGT
TON    AAHCYSN------------NYQVLLGRNNLFK-DEPFAQR        TON
HPHold TAKNLFLN======HSENATAKDIA-PTLTLY===VGKKQL        HPHold
HPHnew TAKNLFLNHSENATAKDIAPTLTLYVGKK==========QL        HPHnew
         --->                  <---->-<->      <-
CON    AAHC                     v λG                    CON
```

Fig. 6. The old[3,4] and new sequence alignments for HPH, relative to those of the known three-dimensional structures, for the region around the SCR 63–71. Note that six residues have been moved from the VR after position 71 to the VR prior to position 63. Nomenclature and labeling are as in Figure 4.

Fig. 4. Sequence alignment for many of the known serine protease sequences. The source of these various sequences and the definitions of the protein name codes used in this Figure are given in Table II. The first seven proteins (above the line) are those with known experimental three-dimensional structures (see Table I). Their sequences have been aligned based on the superposition of the three-dimensional structures as described in the text. The remaining proteins were aligned using the characteristic sequence matching patterns (see text). When the sequence alignment is uncertain, the respective sequence is shown in italics. The IUPAC-IUB convention standard single letter amino acid code, used in this figure, is as follows: A = Ala, C = Cys, D = Asp, E = Glu, F = Phe, G = Gly, H = His, I = Ile, K = Lys, L = Leu, M = Met, N = Asn, P = Pro, Q = Gln, R = Arg, S = Ser, T = Thr, V = Val, W = Trp, Y = Tyr. An asterisk indicates that the sequence continues past the last residue shown in the figure. Positions of relative deletions in the sequences are denoted by a dash in the known structures and by a double dash in the sequences aligned using the characteristic homology patterns. The line labeled "CNO" gives the chymotrypsinogen residue numbering used throughout this paper. The ↔ symbols delineate the SCRs. The "CON" line lists the conserved, characteristic sequence patterns used to align the "new" sequences (see text). They are coded as follows: upper case, almost completely conserved side chain; lower case, high frequency of this amino acid at this position; λ, nonpolar residues, typically A, V, L, I, M; π, denotes a polar residue, typically S, T, Q, N plus the charged residues; o, denotes S or T, (may be substituted by A occasionally); ϕ, denotes aromatic, usually Y or F, sometimes W; +, denotes a positive residue such as R, K, or H; −, denotes a negative residue such as D or E. The line labeled "S–S" gives the position of the half cystine residue that is disulfide bridged to the half cystine at the label position. A question mark is placed at positions where it is not clear to which residue a disulfide bridge is formed.

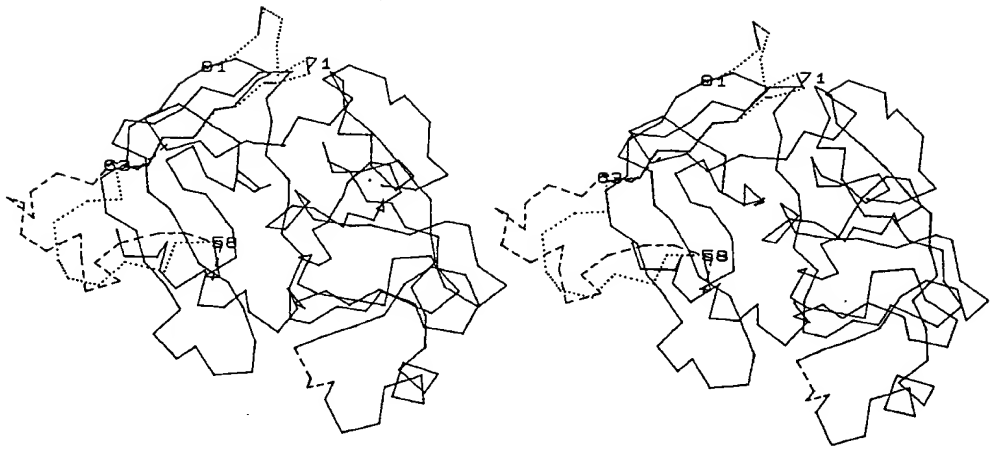


Fig. 7. Cα plots for the old (dotted lines) and new (solid lines) structures of HPH based on the respective alignments shown in Figure 6. The differences appear around the SCR at positions 63–71. Because the VR from 59 to 62 in the new HPH alignment is six residues longer than in the previous alignment (Fig. 6) and at least eight residues longer than any of the currently known structures in this loop (see Figs. 4 and 6, Table III), the conformation that has been drawn for this VR (dashed lines) is arbitrary. It has been included only to give a sense of the relatively large size of this loop. The difficulties associated with finding the correct conformation for such large additions are described in the text. The VR at residues 72–80 is highly truncated, seven residues shorter, in the new structure. Compare the conformations of the known structures for the VRs at 59–62 and 72–80 (Fig. 3) with these loops in HPH. Whereas the Cα positions (and main chain coordinates) for residues 63–71 are the same in this figure for the old and new structures of HPH, the side chains at each residue position are, of course, different in the two alignments (Fig. 6), giving different overall structures for this part of the molecule as well.

Alignment of "New" Serine Protease Sequences

Identification of the characteristic sequence conservation patterns for each SCR allows the process of aligning "new" sequences to begin. Figure 4 shows the alignment of a large number of serine protease sequences to those of the known structures. The proteins included are taken from a very wide variety of functions and species. It is worth noting that three of these proteins, although clearly homologous to the rest and therefore full members of the family, are no longer functional serine proteases. These are the heavy chain of haptoglobin (HPH), protein Z (PRZ), and the α-chain of nerve growth factor (NGA). For all these proteins, the characteristic pattern for each SCR was located and matched. Then the remaining positions in that SCR were filled, without permitting any additions or deletions.

One of the results that emerges immediately from this alignment (Fig. 4) is the strong reinforcement of the characteristic patterns in all these "new" sequences. The same basic patterns appear in virtually every protein, yet, in almost every pattern, there are individual protein sequences that have exceptions. Usually the exceptions are minor deviations from the theme of the conservation. For example, the chosen four-residue stretch from 139 to 142, S/T/A-G-W-G, is present in 26 of 35 protein sequences reported in Figure 4. The Trp is replaced by a Phe or a Tyr, both of which are typically acceptable replacements for a Trp, in four other proteins. In the same way, the characteristic sequence C-G-G almost always appears at positions 42–44, but sometimes one of the Glys becomes an Ala, and occasionally the Cys is replaced by an alternative small side chain. It is clear that, with perhaps the occasional exception of the difficult SCRs between 63–71 and 81–96 discussed above, all the sequences shown can be aligned relatively trivially and unambiguously to the respective characteristic conserved sequences in the SCRs (Fig. 4).

The close examination of the SCRs at 63–71 and 81–96 described in the previous section (Fig. 5), together with the strong reinforcement of the characteristic sequence patterns in these SCRs (Fig. 4), has led to the realignment of the HPH sequence in this region. This results in a large change in the three-dimensional model structure for this protein from that previously proposed.[3,4] The old alignment (Fig. 6) was forced by the need to avoid placing the charged Asp 65 in position 66, where it would be buried in the hydrophobic core of the N-terminal β-barrel. Pro was placed at position 69 as a replacement for a Gly that would permit a turn. Realization that the aliphatic residues at positions 66 and 68 and the almost complete conservation of a Gly at position 69 (Figs. 4 and 5) make up the characteristic sequence pattern for this SCR led to the realignment of the HPH sequence relative to those of the known structures as shown in Figure 6. As a result

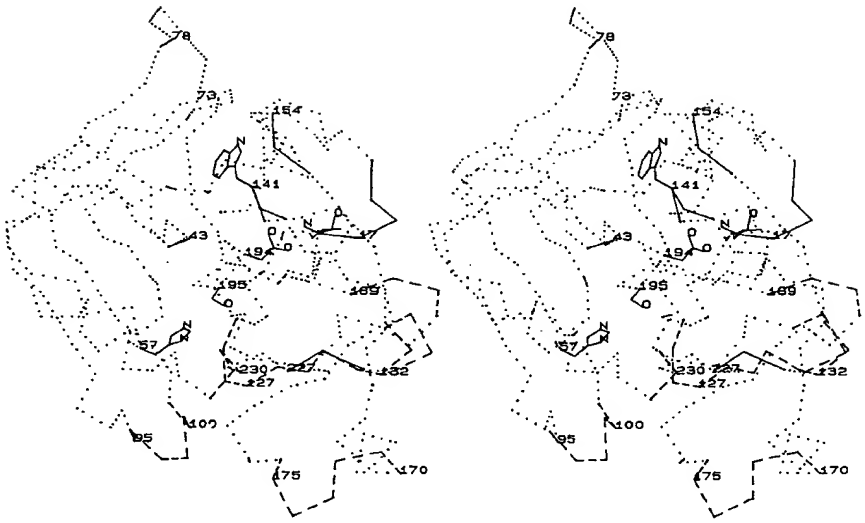


Fig. 8. Cα plot for CHT is shown indicating the sites of VRs with large additions (dashed lines) and with large deletions (dotted lines) in CFB. The unusual residues described in the text are represented by solid lines. These residues include 16–20, 43, 139–142, and 154–156. As can be seen, the unusual residues are localized in one region of the molecule around the normal site of the N-terminus of the protease peptide chain, residue 16. Residues 16 and 17 should not be in their usual positions, because 16 is not N-terminal in CFB or in CF2. Consequently, they really should be placed in the approximate position expected from our knowledge of the zymogen structures[33–36] where the cleavage at 16 has not yet occurred.

of the new alignment, the HPH sequence has a very large addition in the VR at 59–62 and a very short loop in the VR at 72–80. Now, however, the requirements of aliphatic residues at 66 and 68 and Gly at position 69 are completely satisfied. The difference between the old and new structures for HPH is illustrated in Figure 7, which demonstrates clearly that changes in the alignment can result in large changes to the derived structures.

From an evolutionary perspective, deviation from the typical pattern of sequence conservation occurs most often in those members of the family that are no longer serine proteases in function. For example, HPH and PRZ are no longer serine proteases in function. Both have lost the active site Ser 195 (it is an Ala), and His 57 is also replaced by a Lys in HPH. Nevertheless, both proteins retain most of the characteristic sequence homology patterns, yet they deviate more than the typical sequence. Examples of this are the immediate area around the Ser 195 or the SCR at 63–71. The *Streptomyces griseus* trypsin-like protein also has a weaker homology pattern than the rest, as has been previously noted.[4] Even further diverged are the other bacterial serine proteases.[20–22] These latter are not considered in this paper because of the considerably greater difficulty in predicting their structures correctly by these methods (see footnote * on page 318).

Analysis of the Variable Regions

Once the sequences that correspond to the SCRs are identified and aligned, attention can be focussed on the remaining residues forming the VRs. Based on the alignment in Figure 4, the sequence lengths deduced for the various VRs are summarized in Table III. It is clear from Table III that an amazing degree of variation occurs in the length of these loops and therefore in their conformations (Fig. 3).

In previous studies,[4] we classified the modeling of the VRs into five different classes. 1) A known structure has the same-length VR and the same residue character. In this case, a model structure is constructed for the "new" VR by clipping in the respective VR from the known structure. In such cases, the confidence level in this part of the structure can be reasonably good, especially if it is a short VR. One example occurs among the known structures of the serine proteases at residues 97–101, where the VRs of chymotrypsin (CHT), trypsin (TRP), and the mast cell protease (MCP) are all the same length and have the same conformation (see this loop in Fig. 3). 2) Different lengths appear in this VR among the known structures, but they all have the same structural motif, e.g., a β-bend connecting two antiparallel β-strands. The application of a common structural motif to the "new" VR conformation can be illustrated by examining the VR at positions 35–41

the DSN protein in the SCR from 63 to 71. The usual Gly residue that invariably appears at position 69 is absent. This leads to an ambiguity regarding how to align the sequence. However, at least one of the possibilities, the one shown in Figure 4, is likely to fit into the usual serine protease structure at this position with only minor deviations.

On the other hand, there are examples of larger changes, including the complete absence of one or more characteristic sequence patterns. This can be illustrated by the SCR at 134–145. This SCR has the highly conserved, very characteristic S/T/A-G-W-G sequence at 139–142, as noted above. Virtually every serine protease has this sequence, with a rare change from Trp to Tyr or Phe and a very rare change of the second Gly to something else (see PRZ for an example of divergence as discussed in the previous paragraphs). Note, however, the sequence for this SCR in CFB and CF2 (Fig. 4). In these cases, the characteristic residues are completely missing. Clearly something is happening to the structure here that is unusual. The absence of these residues creates a serious problem in properly aligning these two sequences in this SCR. There is no obvious pattern of residues that would correspond to the S/T/A-G-W-G sequence at 139–142 among the residues that are in this region of the sequence. It is possible, in principle, to place the sequence A-L/H-F-V at these positions, thereby replacing the conserved Trp with a Phe. However, this leaves internal positions 136 and especially 138 with charged residues such as Asp and Lys, where nonpolar residues are characteristic and are called for by the structure. The sequences have been tentatively aligned as shown in Figure 4 to satisfy the requirements that internal residues be nonpolar, or at least uncharged. The uncertain alignments are shown in italics in Figure 4.

If we look further at these two sequences, CFB and CF2, we find another anomalous site. The classic, characteristic pattern for positions 42–44 is C-G-G. Examination of the sequences in Figure 4 shows that substitution of one of the Gly residues by Ala occurs occasionally. The startling introduction of a much larger Met side chain at position 43 in CFB and an even larger and charged Arg in CF2 indicates that once again something strange is happening in these proteins. Examination of the location of these changes in the serine protease three-dimensional structure (Fig. 8) shows that they are immediately adjacent to each other in the structure even though they are quite distant in the sequence.

A further anomalous sequence occurs at positions 16–19 of these two sequences. Activation of a serine protease from its zymogen protein involves the cleavage of the peptide chain prior to residue 16 to generate a free amino terminus at this point in the molecule.[33,52] The earliest serine protease crystal structure solutions[24] showed that this amino terminus formed an internal salt bridge with the side chain of Asp 194. The homology observed in the residues at positions 16–19 is likely to be diagnostic of this activation step in all these enzymes. Thus it is interesting that several of the sequences in Figure 4 show major deviations from the characteristic sequence pattern for these four residues. All the serine proteases have Ile or Val (or Leu) at positions 16 and 17 and usually Gly at 18 and 19. Exceptions to this pattern are found in PRZ and NGA, neither of which is a true protease and thus presumably no longer retains an activation step. The two proteins CFB and CF2 are also striking in the absence of any homology to the characteristic pattern for this stretch of residues or even between the two themselves. It is known that, unlike the other serine proteases, these two proteins are clipped some 200 residues N-terminal to residue 16.[53,54] Thus the activation mechanism for these proteins must be different from that of the other serine proteases. Figure 8 shows that the normal location of the N-terminal residues 16 and 17 is also close to the two unusual sequences described above for the three-dimensional structure.

Thus significant violations of three spatially adjacent characteristic sequence patterns appear in these two proteins, coupled with functional significance expressed in the absence of the normal mechanism of zymogen activation. It is interesting to note that CFB and CF2 occupy parallel functional roles in the alternative and classical pathways of the complement system, respectively.[55] They form the proteolytic subunits of the respective C3 convertases in the two pathways and, upon binding of C3b subunits, change their specificity to become C5 convertases. Taken together, these observations suggest that some unusual three-dimensional structure is occurring in this particular region of the molecule involving the activation site and the immediately adjacent active and specificity sites. What is difficult to predict or determine is just how different these structures actually are from the typical serine protease structural theme in this region. The excellent preservation of the characteristic sequence homology patterns everywhere else in the CFB and CF2 sequences (Fig. 4) suggests strongly that the remaining parts of the molecule are close to their normal conformation. However, there is no such reassurance for this part of the molecule.

As one begins to construct this part of the CFB molecule onto the typical SCR framework of the serine proteases, several problems are encountered very rapidly. The introduction of side chains at positions 140 and 43 in place of Gly, in such close proximity, causes a steric collision of these two groups. In constructing the N-terminal portion of the chain, the conformation for residues 16–19 must be taken from the zymogen structure of chymotrypsinogen,[33,34] since these residues cannot reside in the usual N-terminal Ile 16 binding pocket. This leaves the problem of what to do with Asp 194, the other

half of the normal salt bridge to the amino terminus at Ile 16. Both possible conformations of the aspartate, that of the zymogen and of the mature enzyme, must be considered based on zymogen crystal studies.[33-36] Unfortunately, neither conformation seems possible without a significant change in the local CFB structure. In the zymogen conformation, the side chains of Met 43 and Ala 140 (both usually glycines) prevent the Asp 194 side chain from occupying its usual place. In the mature enzyme form, it is the side chain of Phe 142 (also typically a glycine) that is too close to permit the usual conformer (see Fig. 8).

The resulting changes that need to be made to accommodate these unusual side chains are difficult to predict. Does the structure undergo a large number of small changes that results in the accommodation of the above groups in close to the normal conformation, or is there a significant change in this portion of the molecule that gives a very different conformation around Asp 194 and residues 139–142? It is not possible to answer this question using the energetics methods currently available. Such problems may have to wait either for experimental structure determinations or for improved understanding of protein structure and energetics. Therefore, it is particularly important that comparative modeling methods, as described in this paper, allow us to recognize such regions of the molecule that are likely to be particularly difficult to construct reliably.

ACKNOWLEDGMENTS

I am indebted to my colleagues Drs. Stan Burt, John Erickson, and Charles Hutchins for many valuable discussions and for critical reading of the manuscript.

REFERENCES

1. Browne, W.J., North, A.C.T., Phillips, D.C., Brew, K., Vanaman, T.C., Hill, R.L. A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. J. Mol. Biol. 42:65–86, 1969.
2. McLachlan, A.D., Shotton, D.M. Structural similarities between α-lytic protease of Myxobacter 495 and elastase. Nature [New Biol.] 229:202–205, 1971.
3. Greer, J. Model for haptoglobin heavy chain based upon structural homology. Proc. Natl. Acad. Sci. USA 77:3393–3397, 1980.
4. Greer, J. Comparative model building of the mammalian serine proteases. J. Mol. Biol. 153:1027–1042, 1981.
5. Feldmann, R.J., Bing, D.H., Potter, M., Mainhart, C., Furie, B., Furie, B.C., Caporale, L.H. On the construction of computer models of proteins by the extension of crystallographic structures. Ann. N.Y. Acad. Sci. 439:12–43, 1985.
6. Greer, J. Protein structure and function by comparative model building. Ann. N.Y. Acad. Sci. 439:44–63, 1985.
7. Blundell, T., Sibanda, B.L., Pearl, L. Three-dimensional structure, specificity and catalytic mechanism of renin. Nature 304:273–275, 1983.
8. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
9. Devereux, J., Haeberli, P., Smithies, O. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12:387–395, 1984.
10. Arcoleo, J.P., Greer, J. Hemoglobin binding and its relationship to the serine protease-like active site of haptoglobin. J. Biol. Chem. 257:10063–10068, 1982.
11. Mollison, K.W., Mandecki, W., Zuiderweg, E.R.P., Fayer, L., Fey, T.A., Krause, R.A., Conway, R.G., Miller, L., Edalji, R.P., Shallcross, M.A., Lane, B., Fox, J.L., Greer, J., Carter, G.W. Identification of receptor-binding residues in the inflammatory complement protein C5a by site-directed mutagenesis. Proc. Natl. Acad. Sci. USA 86:292–296, 1989.
12. Lustbader, J.W., Arcoleo, J.P., Birken, S., Greer, J. Hemoglobin-binding site on haptoglobin probed by selective proteolysis. J. Biol. Chem. 258:1227–1234, 1983.
13. Greer, J. Model of a specific interaction: Salt bridges form between prothrombin and its activating enzyme blood clotting factor Xa. J. Mol. Biol. 153:1043–1053, 1981.
14. Sham, H.L., Bolis, G., Stein, H.H., Fesik, S.W., Marcotte, P.A., Plattner, J.J., Rempel, C.A., Greer, J. Renin inhibitors: design and synthesis of a new class of conformationally restricted analogs of angiotensinogen. J. Med. Chem. 31:284–295, 1987.
15. Murphy, M.E.P., Moult, J., Bleackley, R.C., Gershenfeld, H., Weissman, I.L., James, M.N.G. Comparative molecular model building of two serine proteinases from cytotoxic T lymphocytes. Protein Structure Function Genet 4:190–204, 1988.
16. Delbaere, L.T.J., Brayer, G.D., James, M.N.G. Comparison of the predicted model of α-lytic protease with the X-ray structure. Nature 279:165–168, 1979.
17. Read, R.J., Brayer, G.D., Jurasek, L., James, M.N.G. Critical evaluation of comparative model building of *Streptomyces griseus* trypsin. Biochemistry 23:6570–6575, 1984.
18. Zuiderweg, E.R.P., Henkin, J., Mollison, K.W., Carter, G.W., Greer, J. Comparison of model and nuclear magnetic resonance structures for the human inflammatory protein C5a. Protein Struct Function Genet 3:139–145, 1988.
19. Polgar, L. Serine proteases. In: "Mechanisms of Protease Action." CRC Press, Boca Raton, Florida, 1989: 87–122.
20. Fujinaga, M., Delbaere, L.T.J., Brayer, G.D., James, M.N.G. Refined structure of α-lytic protease at 1.7 Å resolution. Analysis of hydrogen bonding and solvent structure. J. Mol. Biol. 184:479–502, 1985.
21. Moult, J., Sussman, F., James, M.N.G. Electron density calculations as an extension of protein structure refinement. *Streptomyces griseus* protease at 1.5 Å resolution. J. Mol. Biol. 182:555–566, 1985.
22. Read, R.J., Fujinaga, M., Sielecki, A.R., James, M.N.G. Structure of the complex of *Streptomyces griseus* protease B and the third domain of the turkey ovomucoid inhibitor at 1.8 Å resolution. Biochemistry 22:4420–4433, 1983.
23. Read, R.J., James, M.N.G. Refined crystal structure of *Streptomyces griseus* trypsin at 1.7Å resolution. J. Mol. Biol. 200:523–551, 1988.
24. Birktoft, J.J., Blow, D.M. Structure of crystalline α-chymotrypsin. V. The atomic structure of tosyl-α-chymotrypsin at 2 Å resolution. J. Mol. Biol. 68:187–240, 1972.
25. Chamber, J.L., Stroud, R.M. The accuracy of refined protein structures: comparison of two independently refined models of bovine trypsin. Acta Crystallogr. B35:1861–1874, 1979.
26. Sawyer, L., Shotton, D.M., Campbell, J.W., Wendell, P.L., Muirhead, H., Watson, H.C., Diamond, R., Ladner, R.C. The atomic structure of crystalline porcine pancreatic elastase at 2.5 Å resolution: Comparison with the structure of α-chymotrypsin. J. Mol. Biol. 118:137–208, 1978.
27. Bode, W., Chen, Z., Bartels, K., Kutzbach, C., Schmidt, G., Bartunik, H. Refined 2 Å x-ray crystal structure of porcine pancreatic kallikrein A, a specific trypsin-like serine proteinase. Crystallization, structure determination, crystallographic refinement, structure and its comparison with bovine trypsin. J. Mol. Biol. 164:237–282, 1983.
28. Remington, S.J., Woodbury, R.G., Reynolds, R.A., Matthews, B.W., Neurath, H. The structure of rat mast cell protease II at 1.9 Å resolution. Biochemistry 27: 8097–8105, 1988.
29. Fujinaga, M., James, M.N.G. Rat submaxillary gland

serine proteinase, tonin: Structure solution and refinement at 1.8Å resolution. J. Mol. Biol. 195:373–396, 1987.
30. Blevins, R.A., Tulinsky, A. The refinement and the structure of the dimer of α-chymotrypsin at 1.67 Å resolution. J. Biol. Chem. 260:4264–4275, 1985.
31. Cohen, G.H., Silverton, E.W., Davies, D.R. Refined crystal structure of γ-chymotrypsin at 1.9 Å resolution: comparison with other pancreatic serine proteases. J. Mol. Biol. 148:449–479, 1981.
32. Marquart, M., Walter, J., Deisenhofer, J., Bode, W., Huber, R. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. Acta Crystallogr. B39:480–490, 1983.
33. Freer, S.T., Kraut, J., Robertus, J.D., Wright, H.T., Xuong, N.H. Chymotrypsinogen: 2.5 Å crystal structure, comparison with α-chymotrypsin, and implications for zymogen activation. Biochemistry 9:1997–2009, 1970.
34. Wang, D., Bode, W., Huber, R. Bovine chymotrypsinogen A: X-ray crystal structure analysis and refinement of a new crystal form at 1.8 Å resolution. J. Mol. Biol. 185: 595–624, 1985.
35. Walter, J., Steigemann, W., Singh, T.P., Bartunik, H., Bode, W., Huber, R. On the disordered activation domain in trypsinogen. Chemical labelling and low temperature crystallography. Acta Crystallogr. B38:1462–1472, 1982.
36. Kossiakoff, A.A., Chambers, J.L., Kay, L.M., Stroud, R.M. Structure of bovine trypsinogen at 1.9 Å resolution. Biochemistry 16:654–664, 1977.
37. Remington, S.J., Matthews, B.W. A systematic approach to the comparison of protein structures. J. Mol. Biol. 140: 77–99, 1979.
38. Rossmann, M.G., Argos, P. The taxonomy of protein structure. J. Mol. Biol. 109:99–129, 1977.
39. Waterman, M.S. Computer Analysis of Nucleic Acid Sequences. Methods Enzymol. 164:765–793, 1988.
40. Ponder, J.W., Richards, F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–91, 1987.
41. Bruccoleri, R.E., Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26:137–168, 1987.
42. Bruccoleri, R.E., Karplus, M. Chain closure with bond angle variations. Macromolecules 18:2767–2773, 1987.
43. Shenkin, P.S., Yarmush, D.L., Fine, R.M., Wang, H., Levinthal, C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ring-like structures. Biopolymers 26:2053–2085, 1987.
44. Moult, J., James, M.N.G. An aligorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1:146–163, 1986.
45. Burt, S., Greer, J. Search Strategies for Determining the Bioactive Conformers of Peptides and Small Molecules. Ann. Rep. Med. Chem. 23:285–294, 1988.
46. Hagler, A.T., Stern, P.S., Sharon, R., Becker, J.M. and Naider, F. Computer simulation of the conformational properties of oligopeptides: Comparison of theoretical methods and analysis of experimental results. J. Am. Chem. Soc. 101:6842–6852, 1979.
47. Dauber, P., Osguthorpe, D. and Hagler, A.T. Structure, energetics, and dynamics of ligand binding to dihydrofolate reductase. Biochem. Soc. Trans. 10:312–318, 1982.
48. Kraulis, P.J., Jones, T.A. Determination of three-dimensional protein structures from nuclear magnetic resonance data using fragments of known structures. Proteins 2:188–201, 1987.
49. Struthers, R.S., Hagler, A.T., Rivier, J. In: "Conformationally Directed Drug Design: Peptides and Nucleic Acids as Templates or Targets," Vida, J.A., Gordon, M., eds. Washington, D.C.: Am. Chem. Soc., 1984: 239.
50. Chothia, C., Lesk, A., Levitt, M., Amit, A., Mariuzza, R., Phillips, V., Poljak, R. The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. Science 233:755–758, 1986.
51. Fine, R.M., Wang, H., Shenkin, P.S., Yarmush, D.L., Levinthal, C. Predicting antibody hypervariable loop conformations. II. Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. Proteins 1:342–362, 1986.
52. Stroud, R.M., Kossiakoff, A.A., Chambers, J.L. Mechanisms of zymogen activation. Annu. Rev. Biophys. Bioeng. 6:177–193, 1977.
53. Mole, J.E., Anderson, J.K., Davison, E.A., Woods, D.E. Complete primary structure for the zymogen of human complement factor B. J. Biol. Chem. 259:3407–3412, 1984.
54. Bentley, D.R., Porter, R.R. Isolation of cDNA clones for human complement component C2. Proc. Natl. Acad. Sci. USA 81:1212–1215, 1984.
55. Reid, K.B.M., Porter, R.R. The Proteolytic activation systems of complement. Annu. Rev. Biochem. 50:433–464, 1981.